FAULT RECOVERY METHOD AND DEVICE THEREFOR

Patent Number: JP11024849
Publication date: 1999-01-29
Inventor(s): FUJIBAYASHI AKIRA; WATANABE NAOKI
Applicant(s):: HITACHI LTD
Requested Patent: JP11024849
Application Number: JP19970182105 19970708
Priority Number(s):
IPC Classification: G06F3/06 ; G06F3/06 ; G06F12/16 ; G11B19/02 ; G11B19/04
EC Classification:
Equivalents:

Abstract

PROBLEM TO BE SOLVED: To shorten the fault recovery time to restore only the store area of the logically effective data by expanding this area to a file store position or an idle area included in a physical disk controller. SOLUTION: This fault recovery device is provided with a disk controller 101, a host computer 102, a disk device 103, etc. A file system 105 that is managed by an OS(operating system) 104 of the computer 102 performs the file input/output control by means of a table 106 which manages the store positions of files managed by the system 105 and the idle areas of the device 103. The table 106 is stored in the device 103. Then the fault recovery device produces an effective area table of the device 103 and recovers only an effective area of the device 103 by means of the table 106 which is managed by the OS 104 of the computer 102.

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－24849

(43)公開日　平成11年(1999)１月29日

(51)Int.Cl.[6]		識別記号	FI		
Ｇ０６Ｆ	3/06	３０６	Ｇ０６Ｆ	3/06	３０６Ｂ
		５４０			５４０
	12/16	３１０		12/16	３１０Ｑ
Ｇ１１Ｂ	19/02	５０１	Ｇ１１Ｂ	19/02	５０１Ｆ
	19/04	５０１		19/04	５０１Ｄ

審査請求　未請求　請求項の数４　ＯＬ　（全７頁）

(21)出願番号　特願平9－182105

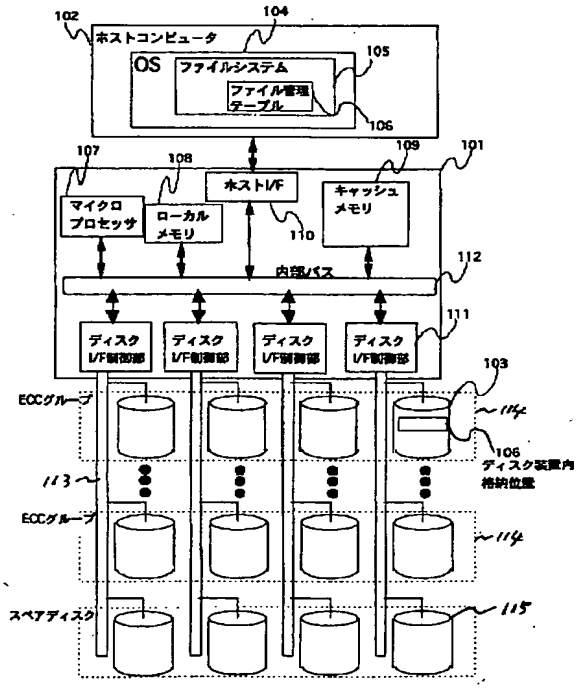
(22)出願日　平成９年(1997)７月８日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　藤林　昭
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内
(72)発明者　渡邊　直企
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内
(74)代理人　弁理士　小川　勝男

(54)【発明の名称】　障害回復方法および装置

(57)【要約】
【課題】ディスク装置の障害回復時間を短縮する。
【解決手段】ディスクアレイ装置において、ディスク装置の障害回復処理時にホストコンピュータのオペレーティングシステム（ＯＳ）の管理するファイル管理テーブルを基に上記ディスク装置の有効領域のみを回復して障害回復時間を短縮する。

図１



【特許請求の範囲】

【請求項１】ディスク制御装置と複数のディスク装置とからなるディスクアレイ装置において、ディスク装置の障害発生時より障害回復処理を行う時に、該ディスクアレイ装置に接続しているホストコンピュータのオペレーティングシステムの管理している論理的なディスク装置内のファイル格納領域または空き領域の状態をディスク制御装置が認識することにより、ディスクアレイ装置内の論理的有効領域のみを回復することを特徴とする障害回復方法。

【請求項２】請求項１記載の論理的なファイル格納領域または空き領域をディスク制御装置が認識する方法として、ホストコンピュータのオペレーティングシステムの管理下にある論理的なファイル格納領域または空き領域を管理するためのテーブルをディスク制御装置に送信するようホストコンピュータに対して要求する手段をディスク制御装置が備え、上記要求を受信したホストコンピュータが上記テーブルの最新の情報を、ホストコンピュータ内のメモリに保存している場合は、ディスク制御装置に対して上記テーブルを送信する手段をホストコンピュータに備えさせることを特徴とするディスクアレイ装置。

【請求項３】請求項１記載の論理的なファイル格納領域または空き領域をディスク制御装置が認識する方法として、ホストコンピュータのオペレーティングシステムの管理下にある論理的なファイル格納領域または空き領域のテーブルの最新の情報を、オペレーティングシステムがディスクアレイ装置内の常に定まった既知の領域に保存している場合には、ディスク装置内の当該格納領域よりディスク制御装置が読み出すことを特徴とするディスク制御装置。

【請求項４】前記論理的なファイル格納領域または空き領域のテーブルを受け取るかまたは読み取ったディスク制御装置が上記テーブルをディスク制御装置内の物理的なディスク装置内のファイルの格納位置または空き領域に展開して、この情報を基にしてディスク制御装置内のデータ格納の管理単位（データストライピングを行っている場合は１つのストライプ列）を一領域として、その領域が論理的に有効なデータの格納領域か空き領域かを示すテーブルを作成することを特徴とするディスクアレイ装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、ディスクアレイ装置において、アクセス不能となったデータの障害回復を行う場合に用いる方法とその方法を用いるディスク制御装置に関する。

【０００２】

【従来の技術】現在ＲＡＩＤ（Redundant Array of Inexpensive Disks）技術を用いて信頼性を高めているディスクアレイ装置では、あるディスク装置に障害が発生し、その格納データにアクセス不能になった場合には、障害ディスク装置と同じ冗長構成グループである他のディスク装置に格納されているデータおよびパリティにより、障害ディスク装置内のデータを復元させる。ここで、ＲＡＩＤ１いわゆるミラーリングの場合は二重化されているもう一方のディスク装置のデータを用いる。一般的には、復元したデータはスペアとしてディスク制御装置に接続しているディスク装置（以下スペアディスク装置）に保存し、スペアディスク装置を障害ディスク装置に代替する。

【０００３】データの復元は、ディスクアレイ装置の稼働中に行われ、ホストコンピュータの障害ディスク装置に対するアクセスは中断させない。従って、ホストコンピュータのアクセス要求がある障害ディスク装置内のデータが優先的に復元され、ディスク制御装置のアイドル時に他の部分が復元され、スペアディスクに格納される。

【０００４】データの復元中には、ディスク制御装置は冗長構成グループ内の障害ディスク装置以外のディスク装置すべてにアクセスをしなければならないため、その処理量は増大し一般的にホストコンピュータに対するアクセス性能は正常時よりも低下する。従って、データの復元に要する時間は可能なかぎり短時間であることが望ましい。

【０００５】従来技術の公知例としては、例えば、ＲＡＩＤを提唱した D. Patterson らの「A Case for Redundant Arrays of Inexpensive Disks（ＲＡＩＤ）エーシーエム　シグモド（ACM SIGMOD）Conference, Chicago, IL, june 1988」やThe RAIDAdvisory Board 発行の「The RAID Book」など、一般的なディスクアレイ装置に関する記事または製品のマニュアル等が挙げられる。

【０００６】

【発明が解決しようとする課題】従来の技術では、障害ディスク装置内のデータをスペアディスク装置に復元するのに要する時間は、ディスク装置の記憶容量に比例して増大する。障害修復時は入出力性能の低下を招くため障害修復時間はできるだけ短い方が望ましい。

【０００７】ここで、ディスク装置内に格納されているデータについて考えると、その中には、ホストコンピュータのオペレーティングシステム（ＯＳ）の持つファイルシステムではすでに無効とされたデータで空き領域となっている場合や未使用の領域等も含まれている。本来これらのデータは復元の必要がない。

【０００８】しかし、従来の技術ではディスク制御装置はファイルシステムに見えている論理的なデータの有効，無効やディスク装置内の空き領域を判断する手段を持たないため、ディスク装置内のすべての記憶領域に対してデータ復元処理をしなければならない。ディスク装置障害時の有効なデータの記憶領域占有率（以下ディス

ク使用率と呼ぶ）が低ければ低いほど無駄なデータの復元を行うことになる。

【０００９】

【課題を解決するための手段】本発明では、障害ディスク装置の修復時、ホストコンピュータ上のＯＳの持つ論理的なファイル格納位置や空き領域を管理するテーブル（以下ファイル管理テーブル）を基にして、ディスク制御装置内で物理的なディスク制御装置内のファイル格納位置や空き領域に展開することで論理的に有効なデータの格納領域のみを復元することで障害回復に要する時間を短縮する。

【００１０】障害回復時は、ファイル管理テーブルに基づいて、ディスク制御装置のデータ格納の管理単位を一領域とした有効領域を示すテーブル(以下有効領域テーブル)を作り、このテーブルにしたがってデータ復元処理を進めることで、有効領域のみ回復し処理時間を短縮する。

【００１１】この管理単位はＲＡＩＤ１のようなパリティを持たない冗長構成のディスクアレイ装置の場合には、ディスク装置のトラック単位としたりシリンダ単位とすることが自由であるが、ＲＡＩＤ３，４，５等のデータのストライプとパリティ格納を行う場合には、パリティを演算するストライプ列を１つの管理単位とする。有効領域テーブルの作成の基となるファイル管理テーブルをディスク制御装置が得る手段として、１）ホストコンピュータにファイル管理テーブルの送信を要求する。２）ファイル管理テーブルの格納位置が既知であるＯＳの場合はディスク装置が読み出す。以上の二通り方法がある。

【００１２】また、有効領域テーブルはビットマップ形式でも良いし、リスト形式で作成しても良い。そして、１）の方法の場合にはファイル管理テーブルをホストコンピュータから受け取るための手段として、ホストコンピュータに対してファイル管理テーブルの送信を要求するコマンドを新たに創設する。また、上記コマンドをホストコンピュータのＯＳが理解できるようにＯＳにもその処理手順をマイクロプログラムに組み込む。

【００１３】

【発明の実施の形態】本発明の提供する障害回復方法と必要な装置を以下に図面を示し実施例を参照して詳細に説明する。

【００１４】図１は本発明に必要なシステム構成の概略である。ディスク制御装置１０１，ホストコンピュータ１０２，ディスク装置１０３の大きく分けて３つの構成要素がある。ホストコンピュータ１０２のオペレーティングシステム（ＯＳ）１０４の管理下にあるファイルシステム１０５ではその管理下のファイルの格納位置やディスク装置内の空き領域を管理するためのテーブル（以下ファイル管理テーブルと呼ぶ）１０６を使用しファイル入出力制御を行う。このテーブルはディスク装置内に格納されている。一方、ディスク制御装置は、マイクロプロセッサ(ＭＰ)１０７，メモリ１０８，キャッシュメモリ１０９，ホストＩ／Ｆ１１０，ディスクＩ／Ｆ制御部１１１，内部バス１１２より構成される。ディスク制御装置とディスク装置はディスクＩ／Ｆ（一般的ＳＣＳＩバス）１１３により接続される。

【００１５】ＥＣＣグループとして、ここではＲＡＩＤ５の場合を例として、４台のディスク装置を１グループ１１４としている。このグループ内の１台のディスク装置に障害が発生した場合には障害発生ディスク装置と同一ＳＣＳＩバス上に接続されているスペアディスク装置１１５に、グループ１１４の他の３台のディスク装置から復元したデータを格納する。

【００１６】本発明では、ホストコンピュータのＯＳの管理下にあるファイル管理テーブルを利用して、ディスク装置内の有効領域テーブルを作成してデータ回復を行う。このファイル管理テーブルをディスク制御装置に認識させるには、１）ホストコンピュータから受信する方法、または、２）ファイル管理テーブルの最新データのディスク装置内の格納位置が常に定まった位置で、そのデータ形式も既知の場合には、ディスク制御装置がその格納位置から読み出すという二通りの方法が考えられる。ここでは、より一般的に、ファイル管理テーブルに関する情報はホストコンピュータのＯＳのみが理解している場合を考えて、上記１）の方法で説明を進める。

【００１７】図２は本発明による障害回復処理時のディスク制御装置の動作のフローチャートである。ここでは、ディスク装置の障害発生を検知し、障害回復処理を開始する。ステップ２０１ではホストコンピュータの送信してくるファイル管理テーブルを格納する為のキャッシュメモリ領域の確保を行う。ステップ２０２では、ファイル管理テーブルの送信要求コマンドをホストコンピュータに発行する。ステップ２０３ではホストコンピュータのファイル管理テーブル送信コマンドを受信する。ステップ２０４では、用意しておいたキャッシュメモリの領域にファイル管理テーブルのデータを格納する。ステップ２０５では、キャッシュメモリより、ローカルメモリに上記テーブルのデータを読み込む。ステップ２０６で有効領域テーブル作成処理を行う。ステップ２０７では有効領域テーブルに従って障害回復処理を実行する。

【００１８】一方、上記２）の方法の場合、ホストコンピュータと通信することなしに、当該テーブルのデータをディスク制御装置が読み出し、キャッシュメモリに格納後、前述のステップ２０５以降の動作を行う。

【００１９】具体的なホストコンピュータとディスク制御装置間の上記テーブルの送受信に用いる方法としては、ホストＩ／ＦがＳＣＳＩである場合を例にとるとディスク制御装置より、イデンティファイ（Identify）メッセージを発行し、それに対してホストコンピュータよ

り、リクエスト　センス(Request Sense) コマンドを発行する。ディスク制御装置の障害回復処理を認知したホストコンピュータはファイル管理テーブルをデータとしてライト(write) コマンドを発行する。この時、write コマンドCDBコントロールバイト部のベンダ固有ビット(ビット7，6)に1を立てて、ファイル管理テーブルデータを書き込むことを明示する。これを受けたディスク制御装置はコマンド解析後、受信データを予め確保しておいたキャッシュ領域に格納する。

【0020】図3はディスク制御装置の有効領域テーブル作成処理のフローチャートである。ステップ301でファイル管理テーブルの格納位置先頭論理ブロックアドレスおよびデータ長または未使用領域の先頭論理ブロックアドレスおよびブロック長を参照する。ステップ302でディスク制御装置の持つディスク装置の論理アドレスと前ステップで参照したファイル管理テーブルの情報を比較する。ステップ303で、前ステップの比較結果に従ってファイルが存在する領域またはファイルの存在しない未使用領域を有効または無効と判断し、実際のディスク装置の格納位置に対する有効領域テーブルを作成する。

【0021】この時、そのディスクアレイ装置が用いているRAID方式により有効領域テーブルの管理単位は異なる。RAID0方式およびRAID1方式ではパリティを用いないので、自由な管理単位で良い。しかし、データストライピングとパリティを用いるRAID3，4，5方式等の場合は、パリティを演算しているストライピング列を管理単位としてテーブルを作成する。ステップ304では完成した有効領域テーブルを基にディスク制御装置が障害回復処理を開始する。

【0022】図4は障害回復処理のフローチャートである。従来の回復処理のパスと本発明のパスを示した。本発明の従来方法との違いは有効領域テーブルに従って有効領域のみを処理して行くことである。

【0023】ステップ401では、先頭の領域から有効領域テーブルを参照して有効なら処理を続け、無効なら次の領域の処理に移る。

【0024】ステップ402では障害時の管理単位を1領域として、障害発生ECCグループ内の正常ディスク装置からこの領域を読み出す。RAID1方式であれば正常なディスク装置からこの領域を読み出す。RAID3，4，5方式の場合は障害発生ディスク装置と同一ECCグループを構成していた残りのディスク装置から、ストライプ列を単位としてこの領域を読み出す。

【0025】ステップ403では、読み出した領域のデータの排他的論理和を演算してこの領域の障害ディスク装置のデータを回復する。ステップ404では、回復した領域のデータをスペアディスクに書き込む。ステップ405は全領域の回復を完了したかどうかの判定である。このステップ401～405の処理を障害ディスク装置内の全領域のデータを回復するまで繰り返す。

【0026】また、障害回復処理中に生じる書き込み要求に対しては、キャッシュメモリにデータを格納後、当該データの格納位置に対応する領域の回復処理が終了するまで、ディスク装置に対する書き込みを保留しておく。

【0027】図5はファイル管理テーブルと有効領域テーブルの変換例を示す。ここでは一例として、ファイル管理テーブル501は論理ボリューム511内に格納されるファイルのファイル名502とそのファイルの先頭論理ブロックアドレス503，データ長504で構成されている、またディスク制御装置の持つ論理ブロックアドレスとディスク装置内ブロックアドレスの対応表505は論理ボリューム番号512，論理ブロックアドレス508，ディスク装置番号509，ディスク装置内ブロックアドレス510より構成される場合を例に取って説明する。

【0028】障害回復時にはディスク装置内を論理ブロックアドレスの0番地から最終番地までをディスクアレイ装置が使用しているRAID方式に合わせて適当な領域（Region）507に区切り、これを回復処理の単位とする。この時ファイル管理テーブルの論理ボリューム番号，ファイルの先頭論理ブロックアドレス，データ長を、実際のディスク装置のアドレスに変換し、このアドレス範囲を含む領域は有効として、有効領域テーブル506で有効／無効のビット508を立てる。ここでは有効領域テーブル506はビットマップ形式にしているがリスト形式としても本発明の効果は変わらない。

【0029】また、ファイル管理テーブルの形式はOSにより各種の形式があるが、論理ボリューム内の空き領域を管理している形式のテーブルの場合は、図5で説明した回復処理の領域にマッピングさせて、その領域を無効領域とするテーブルを作成すればよい。

【0030】ディスク制御装置で有効領域テーブルを作成する方法の場合には、ホストコンピュータからファイル管理テーブルを受信する場合および既知の格納位置からディスク制御装置が読み出す場合のどちらも、キャッシュメモリ109にファイル管理テーブルを格納した後、メモリ108に上記テーブルを読み込み、論理アドレスと物理アドレスとの変換を行い有効領域テーブルを作成する。

【0031】図6は本発明におけるデータおよび制御の流れを図示したものである。図中ではディスク装置601が障害を起こした例を考える。ファイル管理テーブル106は、手順602よりディスク制御装置から要求を受けた、ホストコンピュータより、手順603でディスク制御装置へ送信されキャッシュメモリに一時格納される。手順604でキャッシュメモリからローカルメモリ上にファイル管理テーブル106を読み込み、手順605で有効領域テーブルの作成処理を行った後、手順60

６で、ディスク装置に障害処理手順に従って、読み込みコマンドを発行する。

【００３２】

【発明の効果】本発明により、障害ディスク装置内の有効な領域のみを回復処理することで、ディスクアレイ装置での障害回復に要する時間を短縮できる。ディスク装置内の有効な領域の全領域に対する比率が少なければ少ないほど、本発明の効果は大きい。

【図面の簡単な説明】

【図１】本発明を利用するシステムの概要を示すブロック図。

【図２】本発明におけるディスク制御装置の障害回復処理の一例を示すフロー図。

【図３】本発明における有効領域テーブル作成処理の一例を示すフロー図。

【図４】従来と本発明のディスク制御装置の障害回復処理を示すフロー図。

【図５】ファイル管理テーブルと有効領域テーブルの一例を示した説明図。

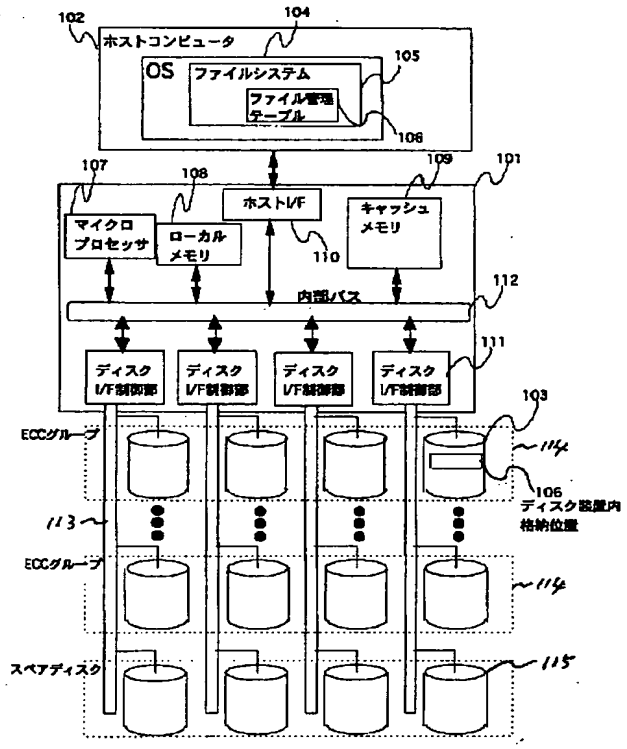
【図６】本発明におけるデータおよび制御の流れを示した説明図。

【符号の説明】

１０１…ディスク制御装置、１０２…ホストコンピュータ、１０３…ディスク装置、１０４…オペレーティングシステム（ＯＳ）１０５…ファイルシステム、１０６…ファイル管理テーブル、１０７…マイクロプロセッサ（ＭＰ）、１０８…メモリ、１０９…キャッシュメモリ、１１０…ホストＩ／Ｆ、１１１…ディスクＩ／Ｆ制御部、１１２…内部バス、１１３…ディスクＩ／Ｆ、１１４…ディスク装置グループ、１１５…スペアディスク装置。

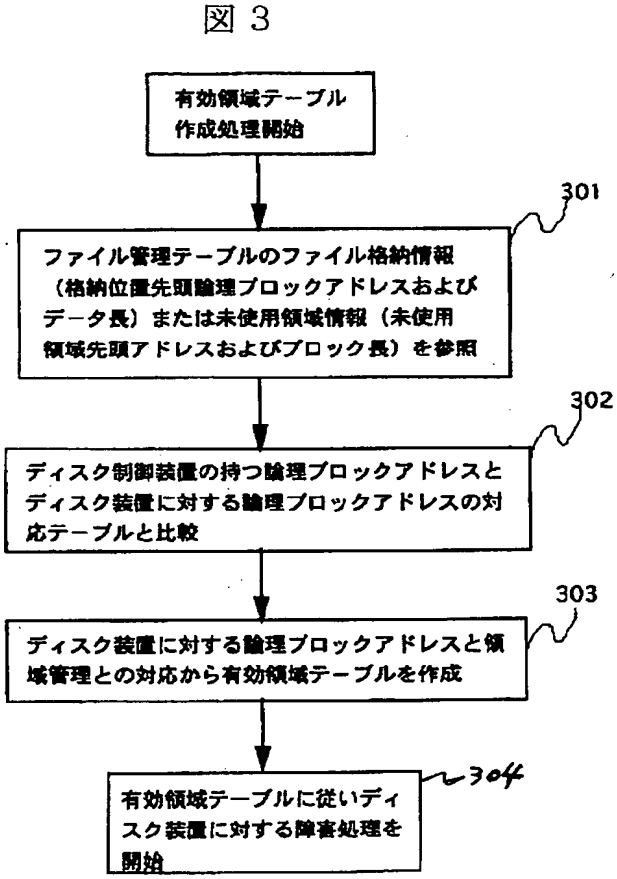
【図１】

図１



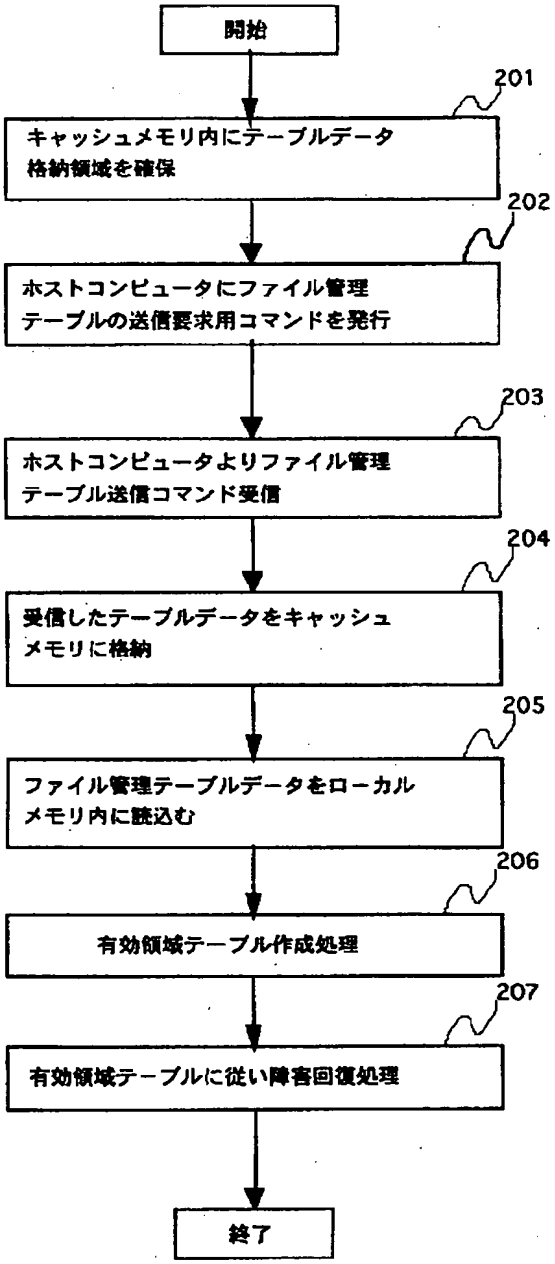
【図３】

図３



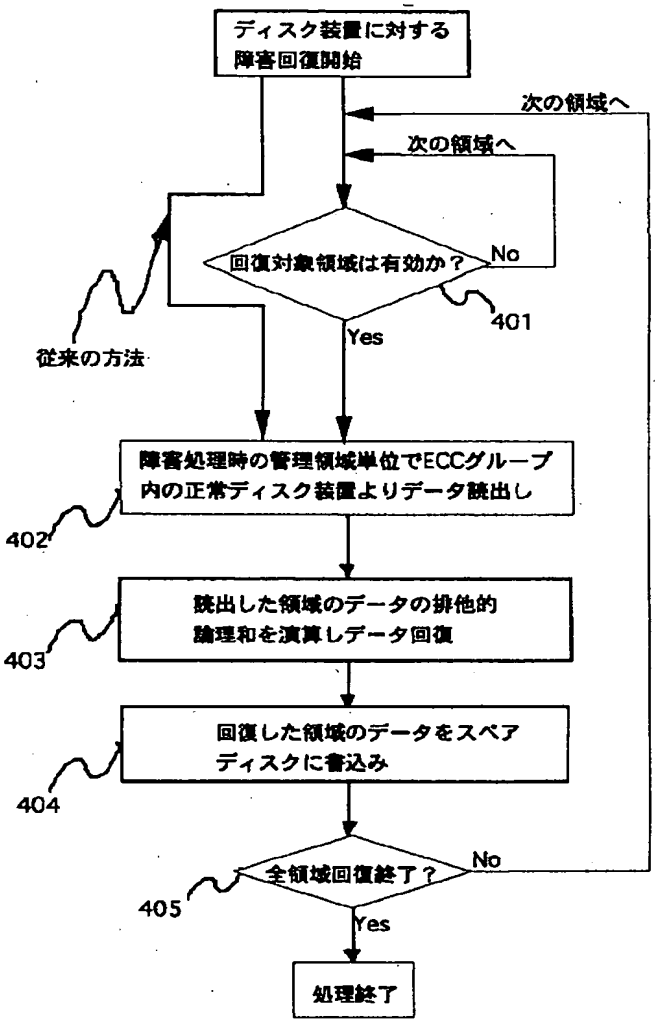
【図２】

図 2


開始
201
キャッシュメモリ内にテーブルデータ格納領域を確保
202
ホストコンピュータにファイル管理テーブルの送信要求用コマンドを発行
203
ホストコンピュータよりファイル管理テーブル送信コマンド受信
204
受信したテーブルデータをキャッシュメモリに格納
205
ファイル管理テーブルデータをローカルメモリ内に読込む
206
有効領域テーブル作成処理
207
有効領域テーブルに従い障害回復処理
終了

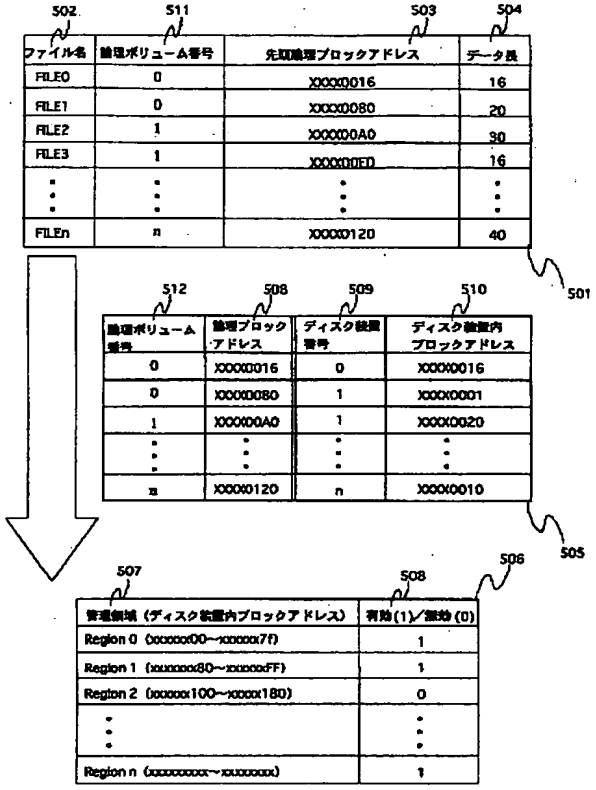
【図４】

図 4


ディスク装置に対する障害回復開始
次の領域へ
次の領域へ
回復対象領域は有効か？
No
401
Yes
従来の方法
障害処理時の管理領域単位でECCグループ内の正常ディスク装置よりデータ読出し
402
読出した領域のデータの排他的論理和を演算しデータ回復
403
回復した領域のデータをスペアディスクに書込み
404
全領域回復終了？
No
405
Yes
処理終了

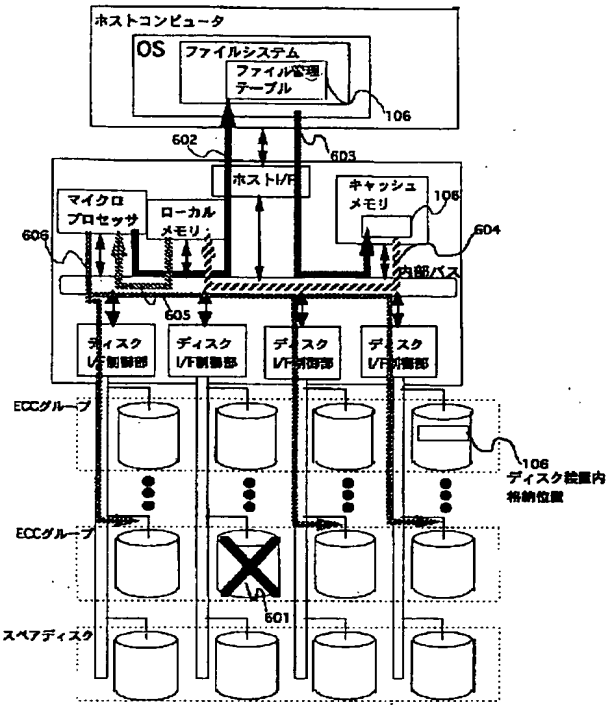
【図５】

図５


502
511
503
504
ファイル名
論理ボリューム番号
先頭論理ブロックアドレス
データ長
FILE0
0
XXXX0016
16
FILE1
0
XXXX0080
20
FILE2
1
XXXX00A0
30
FILE3
1
XXXX00E0
16
FILEn
n
XXXX0120
40
501
512
508
509
510
論理ボリューム番号
論理ブロックアドレス
ディスク装置番号
ディスク装置内ブロックアドレス
0
XXXX0016
0
XXXX0016
0
XXXX0080
1
XXXX0001
1
XXXX00A0
1
XXXX0020
n
XXXX0120
n
XXXX0010
505
507
508
506
管理領域（ディスク装置内ブロックアドレス）
有効(1)／無効(0)
Region 0 (xxxxxx00～xxxxxx7f)
1
Region 1 (xxxxxx80～xxxxxxFF)
1
Region 2 (xxxxxx100～xxxxx180)
0
Region n (xxxxxxxx～xxxxxxxx)
1

【図６】

図６


ホストコンピュータ
OS
ファイルシステム
ファイル管理テーブル
106
602
603
ホストI/F
キャッシュメモリ
108
マイクロプロセッサ
ローカルメモリ
606
604
内部バス
605
ディスクI/F制御部
ディスクI/F制御部
ディスクI/F制御部
ディスクI/F制御部
ECCグループ
106
ディスク装置内格納位置
ECCグループ
601
スペアディスク